学校编码：10384　　分类号_____密级_____
学号：X2013231295　　UDC_____

厦门大学

工 程 硕 士 学 位 论 文

基于自然语言处理的观点句抽取系统设计

Design of the Objective Sentence Extraction System based on Natural Language Processing

李路

指导教师姓名：　赖永炫　副教授
专 业 名 称：　软 件 工 程

论文提交日期：　2016 年 3 月
论文答辩日期：　2016 年 5 月
学位授予日期：　2016 年 6 月

指 导 教 师：_____
答辩委员会主席：_____

2016 年 3 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果,均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行)》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名)：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

互联网的快速发展，特别是 Web2.0 的概念和技术的应用与推广，带来了全新的媒介形式、社群环境和营销理念。在这种环境下，以互联网为媒介的评论、观点和意见等主观性文本信息成指数级增长，文本意见挖掘技术逐渐成为语言信息处理领域的研究热点。其中如何抽取观点句，即将主观性评论句与客观性描述句区分开来，是文本意见挖掘技术中基础且重要的一环。它不仅可以让用户更快捷地找到产品的相关评价内容，也可以让生产厂家对此产品得到及时的反馈从而进行更深入的研究改进。

本文针对目前观点句抽取领域的现状，重点研究了如何从互联网特定领域的非结构化文本中获取相关信息并抽取观点句的技术，并构建了一个基于自然语言处理的观点句抽取系统。本文的主要工作包括：

1.本文提出了一种融合链接密度与内容相似度的网页正文提取算法。该方法与以往使用的 DOM 树解析方法不同，无需使用网页分析工具，而是将网页源码看作一个字符串，并按标签分割为不同的节点，利用节点内容与标题内容的相似度以及节点内的链接密度来判断其是否为正文的起始或结束节点。

2. 本文提出了一种 SVM 和组合列表密度相结合的网页评论提取算法，即新闻正文下面大多数都包含诸如：“有 XX 人参与评论”、“本文共有 XX 条评论”的超链接文字，识别提取此类超链接并重定向，可获取短文本候选网页。

3. 本文提出了一种规则与统计、粗粒度与细粒度相结合的观点句抽取方法。粗粒度提取中，融合观点特征词、句法特征和依存特征进行提取；细粒度提取中，设计一种全新的 CSR 序列提取算法，并结合语义角色信息与 CRF 条件随机场进行提取；最后选取不同的特征组合，利用支持向量机 SVM 分类器，完成观点句的抽取工作。

关键词：观点句；正文提取；评论提取

ABSTRACT

With the rapid development of World Wide Web (WWW), especially the application and promotion of its concept and techniques, Web 2.0 has brought a brand new form of media, a concentrated social environment and an innovation in marketing philosophy. The extraction of opinionated sentence is the basic and important procedure of opinion mining. Its main task is to distinguish the subjective sentence from the objective sentence. By using this technique, the users can easily find the evaluation of the products. Besides it's convenient for the manufacturers to carry on the market research.

This paper is focus on the extraction technique from the domain specific and unstructured information of the Internet. The main research work and contributions are listed as follows:

1. We propose an algorithm for content extraction by using the link density and node weight. This algorithm does not depend on DOM tree, and does not need other web parser. The web page source code is seen as a string, then it is divided into different nodes by labels. With the similarity of the node content and the title and the link density, we can determine whether it is the content node or not.

2. This paper proposed an algorithm of identifying and extracting the review text based on rules and statistics' methods.After analyzing a large number of web pages, we found that there are many hyperlinks below the content of news pages, such as. If we can identify and extract such links, and then redirect, web page contained review text will be obtained.

3. This paper described how to recognize objective sentences from Chinese text documents by applying rule-based methods and statistical methods as well as analyzing the performance of these methods. This method got the broad extraction results by using a lexicon-based method, sentence structure and dependent relationship analysis. Then a kind

of CSR rule extraction algorithm was designed to extract the sequential features. The paper also used a CRF algorithm to identify entities and semantic roles. Finally, the paper found the most optimum and effective features combination to finish the accurate extraction by using SVM classifier and choosing distinguished feature dimensions.

Key words： Objective Sentence; Content Extraction; Review Extraction

目录

Contents

第 1 章 绪论

1.1 研究背景及意义

近年来，随着计算机和网络技术的迅猛发展，互联网用户数量也随之得到了迅速的提升，继而在以报纸、电视、广播为首的传统新闻媒介发展之后，互联网成为了新时代信息提供的主力战场。目前，互联网用户主要通过传统网页、微博、微信、论坛BBS、博客、搜索引擎、RSS 聚合新闻等访问网络信息。现代网络信息为用户提供了一个交流信息的平台。情感倾向分析任务(sentiment orientation analysis)，一般又称为观点抽取(opinion extraction)、观点挖掘(opinion mining)等，目的是分析海量的评论文本中所包含的情感倾向分析。同时，情感倾向分析的研究具有较大的社会和经济价值，可以应用于政治社会科学、商业智能等多个领域。

文本信息主要可以分为客观和主观性文本两种类型。客观性文本信息表达的是客观性陈述；主观性文本往往富有个人观点、情感和态度信息等[1]。而主观性的文本信息的存在的价值是作为其他人做决策时所需参考的他人意见[2]。他人的意见反映了社会群体对于某个事件、人物、产品等对象的倾向性，这种群体智慧可以用来对未来进行准确有力的预测。卡内基梅隆大学在研究了十亿条微博消息后发表论文称，Twitter上主观性挖掘的结果与公共投票的结果高度一致[3]。惠普实验室（HP Lab）的一项研究[4]也表明，可以通过对 Twitter 上的电影评论进行人工的情感分类从而预测电影票房，并且这一应用可以扩展到其他领域，比如产品的销量预测、选举结果预测等。具体而言，主观性文本的重要性首先体现于电子商务领域。对于互联网愈来愈多的在线消费者和经营者来说，产品的评论或者个人使用经验比任何广告都具有说服力和推广价值，消费者更倾向于相信其他消费者的推荐和建议。与此同时，掌握这些评论和主观性信息也是经营者了解市场并取得成功的关键。另一方面，在社会国家安全领域主观性信息也具有及其重要的作用。这些文本信息为互联网舆情监测提供了反映网络社

会群体舆论倾向性的原始数据。总之，主观性文本为我们分析互联网用户行为提供了一种实时的、可理解的途径。

此外情感倾向性分析与文本检索和信息抽取的结合目前已经取得了文本检索会议和亚洲语言信息检索评测会议的关注。观点句抽取是近几年信息处理、自然语言理解领域的一个新的研究方向，已经成为学术界和工业界所关注的焦点。目前，在观点句抽取研究领域研究还并不是很深入，摆在我们面前的仍有很多亟待解决的难题与挑战。因此，对于观点抽取以及观点挖掘方面的技术，不仅具有重要的学术意义，而且具有重要的实用价值。

1.2 国内外研究现状

主观性分析与判别是近几年以来非常热门的研究话题。有许多专家学者采用不同的方法来使用各种各样的应用处理文本从而解决这一问题。尽管大多数人都达到了一定的效果，但是还有很大的发展空间，比如在多语言、跨语言领域以及新型非结构化文本方面需要投入更大的努力。接下来本文简要讨论下当前形势下对于主观性判断的研究成果和现状，以及未来的发展趋势。

观点句抽取任务作为评测会议的一项经典任务，总体上可以分为以下三种方法：即基于情感词典的方法，基于语料库的方法，以及两种方法的结合。

基于情感词典的方法具有领域性，依赖领域相关的情感词典，在新领域语料上的表现不佳。王中卿等[5]将观点句的抽取任务转化为中文文本的主客观分类任务，将只包含一个强情感词、或两个以上弱情感词的文本看作主观文本，利用多部情感词典集合中的词语对中文文本进行主客观分类。李岩等[5]通过统计句子中正负情感词的个数来判断句子的正负倾向性，如果正向情感词大于负向情感词，则将句子判定为正向，小于则为负向，等于则为混合观点句，如果句子中情感词个数为零，则判定为非观点句；朱艳辉等[5]首先将包含观点词的句子作为候选观点句，再通过一些规则的方法对句子进行筛选，获取观点句集合；宋施恩等[5]通过将观点词表和情感词表中的观点词

和情感词设定不同的权值，最后根据句中出现的这两种词进行加权求和，从而获得了观点句。

基于语料库的方法融入机器学习知识，能捕捉到观点句中的隐含特征规律，但由于其需要大量的人工标注语料，所以在现实任务中，往往需要耗费大量的人力在标注上。徐睿峰等[5]利用融合线性分类器和逐步求精的分类器投票的方法，获取最终的观点句集合。从当年评测的结果可以看出，使用这种方法的观点句抽取取得了比较好的结果。由此我们可以得出如下结论，在观点句抽取研究领域或者信息抽取领域，进行实验的数据集将对最终的实验效果产生重大影响，简而言之，如果可以获得的标注的数据集在规模和质量上都有保障的话，可以大大提升整体抽取的效果，反之亦然。北京邮电大学的采用了一种全新的思路，他们将待处理的句子进行句法结构分析然后将句中的所有形容词进行褒贬义的识别并进行训练构建为 CRF 模型，其中每个形容词根据褒贬含义来判断其倾向性，最终利用这一 CRF 模型来判定句中的实体信息，从而获得观点句。

考虑到上文两种方法各自的优势，开始有学者考虑将两种方法进融合，进行观点句抽取。赵立东等[5]综合考虑情感词典、规则和统计机器学习方法下的观点句抽取结果，并对其进行融合，进而得到了观点句集；韩先培[5]同样将句子分为褒义、贬义、中立三类，同时加入了领域信息，在实现上根据情感词词典以及领域特定词分别构建其特定的分类器，其中领域分类器采用了自学习的方法；董喜双等[5]首先利用自建的情感词词典来初步判定句子是否为观点句，然后根据最大熵理论将上述步骤中得到的初步筛选集进行细切分，对每个短句进行情感倾向性判定，最终由短句的判别结果得到每个观点句的情感倾向性，实验结果表明，基于情感的词汇和语料库方法为基础的方法相结合，取得了比较好的结果。

此外，除了评测会议，国内外研究学者也对观点句抽取展开了研究。通常意义上来讲，我们会把句子的含义理解为它词义的集合，因此判定一个句子的情感倾向性首当其冲的就是判定句中的词是否为情感词以及每个情感词的极性[6]。Turney 和

Littman[7]选择七个基础积极和消极的情感词汇作为种子词，来判定其他词的语义倾向，通过考虑上下文信息对种子词进行扩展来研究情感倾向。Sista 等人[8]使用 Word Net 扩展情感词，利用机器学习的方法进行观点句抽取。Hu 与 Liu[9]利用 Word Net 得到形容词种子，以 Bootstrapping 算法拓展情感网络。Hatzivassilolou 等人[10]使用大规模语料信息进行统计，识别出情感词。何婷婷等人[11]基于 How Net 进行相似度计算，识别情感词汇的倾向性。王素格等人[12]对情感词汇同义词的倾向性进行分析，进行情感倾向性判定。

观点句抽取问题可以转化为文本分类问题，可以采用已有的朴素贝叶斯、SVM 等分类算法[13-14]。Pang 等[15]首先将分类的思想进行了实践，在对影片评论的情感倾向性判别研究中，之后又有研究者在此基础上对特征选择进行改进，加入词频、词性标注(POS)、情感词汇和短语、句法关系等特征[16-20]来提高情感分类的结果。Bethard Steven[21]通过训练出的扩展语义句法分类器，标记出句子的观点和观点持有者。Kim，Hovy[22]同样将观点句抽取问题视为句子情感倾向性判定问题，将具有极性的句子视为观点句。Wiebe 等人[23]采用基于文本聚类的方法，首先抽取出每个句子的词频、词组、词性，然后根据这些特征进行观点句的聚类，接着把聚类的到的句子集中的情感词提取出来作为主观词词典构建主观识别模式，并构建分类器完成最终的观点句抽取。Pang 等人[24]在观点句抽取研究方面另辟蹊径，根据最小图分割法将句子中含有主观表达的成分提取出来，并由此进行接下来的识别分类工作，最终完成主观和客观的句子分类的目的。观点句抽取问题除了被视为分类问题外，还被视为回归问题，通过打分实现[25-28]。Qiu 等[29]对知网种子词集合进行迭代扩展，并将第一步得到的集合作为训练集训练支持向量机 SVM，然后再将分类结果反过来迭代用于校正第一步产生的结果。Tan 等[30]采用词汇和语料库相结合的方法，利用情感词汇标注语料训练出一个有监督的分类器进行识别。Melville 等[31]使用贝叶斯分类器，融合情感词汇进行识别。

在情感词的研究方面，除了使用现有的资源如 Word Net、How net 等，比较常用的方法是使用 Bootstrapping 的思想扩展种子集，最终获取一个较大的且与领域相关的情感词汇知识库，此种方法也可应用于观点句抽取中，获取更大更全面的观点句集。